The Journal of Educational Research

ISSN: 0022-0671 (Print) 1940-0675 (Online) Journal homepage: http://www.tandfonline.com/loi/vjer20

A New Kind of School Examination

William A. Mccall

To cite this article: William A. Mccall (1920) A New Kind of School Examination, The Journal of Educational Research, 1:1-2, 33-46, DOI: 10.1080/00220671.1920.10879021

To link to this article: http://dx.doi.org/10.1080/00220671.1920.10879021

A NEW KIND OF SCHOOL EXAMINATION

WILLIAM A. McCALL

Teachers' College, Columbia University, New York

There are in the United States about six hundred thousand school teachers. It is a very conservative estimate that each teacher gives on the average twenty examinations a year. This makes twelve million examinations each year. The time required to construct, give, and score each examination will average, say, three hours. This means that thirty-six million hours are spent examining pupils. The data just given will suffice to show the enormous importance of examinations. Without a doubt, they are and will be for some time and may possibly always remain the most important form of educational measurement. Since this is so, it may seem that those of us who are interested in educational measurement have, in our enthusiasm for constructing and standardizing tests, neglected the traditional type of educational measurement. Really, however, this has not been neglect on our part, for standardized tests are nothing but improved examinations. Furthermore we have been learning new technics which will in time react to improve the making of examinations. The purpose of this article is to show teachers how they may make use of one of these new technics of scientific testing not only to improve certain kinds of examinations but also to make examinations a real pleasure instead of an onerous task to both teacher and pupils.

Below is an illustration of a new way to make an examination. It is designed to test a pupil's knowledge of certain facts concerning the physical features of the United States. We have purposely written it hastily in order that it might illustrate certain crudities of construction. Any teacher in the elementary school could do as well and most teachers could do better.

The examination as presented here assumes that the statements whose truth and falsity are to be determined by the pupils have been mimeographed so that a copy of the examination can be placed in the hands of each pupil. The sample examination given below is supposed to have been worked through by a pupil and then scored by a pupil or by the teacher. The *check, cross*

and *O* mean respectively that the pupil's answer (i. e. his underlining) is correct, incorrect or omitted. Due to limitations of space, only enough of the examination is shown below to illustrate the procedure.

SAMPLE EXAMINATION ON UNITED STATES

Some of the following twenty statements are true and some are false. When the statement is true draw a line under *True;* when it is false draw a line under *False.* Be sure to make a mark for every statement. If you do not know, guess.

1. In general the mountain ranges run east and west	True	False	√
2. Most of the rivers flow north	True	False	√
3. Mt. Mitchell is the highest point east of the Mississippi River	True	False	X
4. Mt. Washington is higher than Mt. Mitchell	True	False	X
5. The Catskill Mountains are in Maine	True	False	√
6. The Cascade Mountains are nearer the Pacific Ocean than the Rocky Mountains..	True	False	X
7. The Rocky Mountains are nearer the Pacific Ocean than the Appalachian Mountains..	True	False	√
8. The Blue Ridge is in the Rocky Mountains.	True	False	√
9. There are more active volcanoes in the west than in the east	True	False	√
10. "Old Faithful" is the name of a cyclone which sweeps upward from Texas into Oklahoma.	True	False	X
11. The "Grand Canyon" was cut through the Cumberland Plateau by the Susquehanna River.	True	False	√
12. Pike's Peak is in the Rocky Mountains.	True	False	√
13. The Mississippi River flows into the Great Lakes.	True	False	√

14. All the following are tributaries of the Mississippi River: Arkansas, Missouri, Ohio... True False V
15. The Big Sandy is the biggest river in the United States... True False X
16. The Atlantic Ocean is to the east and the Pacific Ocean to the west. ..... True False O
17. Canada is to the south and the Gulf of Mexico to the north. True False V
18. The Great Lakes are five in number. ..... True False V
19. It is easier to sink while swimming in the largest lake east than in the largest west of the Mississippi. ..... True False V
20. The central portion of the United States is on the whole more level than the eastern or western portion. ..... True False V

Number of correct underlinings = 14
Number of incorrect underlinings = 5
Number of omissions = 1

Pupil's score = (number correct) — (number wrong). (A)
Pupil's score = 14—5=9

Let us consider first the reason for expressing a pupil's score as the number correct minus the number wrong. Imagine a pupil who is absolutely innocent of any knowledge of the physical features of the United States. Were such a pupil to take the above test and were he to mark every statement, he would according to the theory of chance mark ten statements correctly and ten incorrectly. The chances of his guessing right or wrong are fifty-fifty or one to one. His score on the above test would be:

$$\text{Score} = 10 - 10 = 0.$$

In short the pupil's knowledge is zero and the method of computing his schore gives him zero. Suppose instead that he knows ten statements and guesses at the other ten. Of the ten guessed at he would, according to chance, get five correct and five wrong. That is, even though his real knowledge is ten he will show fifteen

correct (10 + 5) and five incorrect. The method of computing his score brings out his real knowledge.

$$\text{Score} = 15 - 5 = 10.$$

A pupil who marks every statement correctly makes a perfect score, viz:

$$\text{Score} = 20 - 0 = 20.$$

Observe that no account is taken of omissions. Only the corrects and incorrects figure in the pupil's score. When the time allowed the pupils to take the test is made short in order to test each pupil's rate of work there will, of course, be many papers showing several omissions each. In all such cases omissions should be ignored, just as we have done above, in computing scores. Even when the time allowed for the test is ample for each pupil to mark every statement, there will still be an occasional instance of omission due to carelessness, or a misunderstanding of instructions or a puritanic conscience against increasing the score by gamble guess-work even when the instructions urge guessing. When the time is ample for even the slowest pupils and when all are instructed to mark every statement it is much more convenient to compute a pupil's score according to the following formula:

Score = (number of statements)—*2* (number marked incorrectly) . . . . (B).

To derive this formula from the formula marked (A) above, note that since there are no omissions, the number of corrects (C) plus the incorrects (W) equals the number of statements (N). Hence we may write C=N—W. Substituting this in formula (A) gives:

$$\text{Score} = (N-W)-(W)=N-2W.$$

If there are 20 statements in the test and if five are marked incorrectly,

$$\text{Score} = (20) - 2\,(5) = 10$$

Formula (A) gives the same results, i. e.:

$$\text{Score} = 15 - 5 = 10$$

Both formulae give identical results provided there are no omissions. Formula (A) is basic and should be used when there are omissions. Formula (B) should be preferred when there are no omissions or when they are present only in negligible amount. Formula (B) is much more convenient. The

first number is always the same and since the second number is twice the total statements marked incorrectly, it is only necessary to score and total the errors.

It is very difficult for some people to believe that such a test as we have outlined above does anything more than give the highest score to the luckiest guesser. They look with an eye of suspicion upon this thing we call *chance*. Being in a position which offered excellent opportunity, namely treasurer of a Sunday school, we once tossed pennies for heads or tails fifty thousand times. The results came out 25,000 heads and 24,999 tails. Had there not been a miscount somewhere the two would doubtless have come out exactly even. We had occasion to *watch* two summer-school teachers engage in that soul-absorbing, nerve-racking game of chance called *matching pennies*. Each began the summer with a special bag of one hundred pennies. They matched for several minutes daily. The last we heard they were still matching pennies and chance had prevented either from getting complete possession of the other's one hundred pennies. Chance is fatally exact when the pennies or the statements in the test are numerous. The opportunities for injustice in scores multiply in proportion as the number of statements is reduced. Hence there should be as many statements in the test as practical limitations will permit.

The possibility of unreliability of pupil scores is one which this sort of examination shares with all other sorts of tests and examinations. Last summer we gave a test of this kind to our class in educational measurement. As luck would have it two visitors were present and were persuaded to take the examination along with the rest. There were forty statements in the test. The score of one visitor was zero and of the other -2 (when scores are negative they may be considered zero. It is probably the accidents of chance which show an individual to know less than nothing, though a genuine negative score is possible where an individual has been taught errors.) The highest score made by regular members of the class was 36, the lowest score was -4, and the next to the lowest score was 6. The student who made -4 admitted that he was not sure of the truth or falsity of a single statement and that unexpected outside demands had until then

prevented him from studying. Even though he was not given an F at the end of the summer session and even though the course is not required, he appeared at the college this last September and registered for the course again. The student who made a score of 6 registered late. He had been in the course only four or five days. The individual who made the highest score was George Melcher, Director of the Bureau of Research, Kansas City, Mo. All were in agreement that he knew more about educational measurement than anyone else in the class.

During a semester course in educational psychology we gave four examinations. The first was of the *True-False* variety, the second was of the traditional *What* and *Discuss* kind, the third was *True-False,* and the fourth was traditional like the second. Each test lasted about thirty minutes. The half-dozen best and poorest students according to the first test remained with negligible variation the half-dozen best and poorest in all the examinations no matter what the type. An amusing and pathetic incident of the course was the challenge of the poorest students to the rest of the class. Those who had been lowest all through the term proposed in the last test to displace some of those who had been consistently above them. There was a sudden rise in the quality of everyone's work, but the lowest remained the lowest. These words should not be interpreted as meaning that the *True-False* test is unusually reliable. So far as our inadequate data go, there is no reason to suppose that it is much more reliable than the traditional examination.

There are a few suggestions which will help teachers who may wish to use the *True-False* test. First, the teacher should so construct the test that it will contain approximately the same number of true and false statements. A clever pupil may get a higher score than he deserves if he discovers there are many more true statements than false statements in the test or *vice versa.* Suppose there are many more true statements than false statements and suppose some pupil discovers this by observing the statements that he knows, or by observing the teacher's bias for writing true statements instead of false ones. Naturally when he does not know what to mark he will mark *True* thereby securing a little larger score than his ability justifies. It is probably

by just such utilization of the errors of others that the intelligent get through life so much more smoothly than the stupid. On the other hand, the teacher should not have exactly the same number of true and false statements each time, because this will invite clever pupils to count back to see how many more true statements have been marked than false statements. Sometimes there should be more true statements, sometimes more false statements, sometimes the same number of each. A regular plan either as to content or arrangement should be carefully avoided. All the true statements should not come first, neither should the true and false statements be alternated as a regular plan. Let chance determine how many shall be true and how many shall be false and in what order the true and the false shall come.

Second, the teacher should be careful to keep out of the test all ambiguous statements. Statement number 18 in our sample test is somewhat ambiguous. It says: "The great lakes are five in number." Since *great lakes* is not capitalized a pupil might very legitimately interpret this to include Great Salt Lake and others. It will later be difficult to satisfy this pupil that his score should suffer because of the construction he gave this sentence. If the teacher will study her mistakes in this respect she will soon learn how to reduce such ambiguities. As any teacher can testify, the danger of ambiguities of wording are not peculiar to this test. This type of test does not, however, give a pupil an opportunity to reveal just what interpretation he places upon each statement. If the teacher follows the procedure of having pupils score their own or each other's paper all cases of serious ambiguity will be discovered. Statements which are particularly flagrant in this respect can be omitted in scoring.

Third, the teacher should inspect not only this but any sort of test from the point of view of just what the test measures. Statement 19 in our sample test illustrates our point. Our purpose is to test whether the pupil knows that the largest lake west of the Mississippi River contains more salt than the largest lake east of the Mississippi. Instead of measuring this we may be testing whether a pupil knows that it is easier to sink in fresh water than in salt water. Complex wording, unfamiliar terms, the use of

negatives, all tend to make the test a linguistic one. Simple, brief statements without negatives are best.

Fourth, the teacher may so construct the examination as to force pupils to guess wrong due to the power of suggestion. This probably explains why statement 15 was marked wrongly. The pupil doubtless argued to himself that since the river is named the Big Sandy it probably is the biggest river in the United States. The influence of having many suggestive statements in the test is to make the examination more difficult. It operates to give to the pupil who knows nothing at all in the test a large negative score instead of a zero score and it penalizes rather heavily the pupil who does much guessing, for every time he allows himself to be suggested in the wrong direction a point is substracted from the score he has already made by what knowledge he has. In other words, the suggestive statements make the gap between those who know much and those who know little wider than it otherwise would be. Whether a pupil should be specially penalized for allowing himself to be maneuvered wrong is an arguable question. There may be situations where it is eminently desirable to determine whether pupils know what they know so well as to be able to resist suggestion. But we believe that, in general, it is best to avoid suggestive statements. The ideal should be so to construct the examination that any pupil who knows absolutely nothing about the test will make a score of zero.

So much for the construction of the examination. How shall it be applied? The best way, as shown by our sample, is to print, mimeograph, or otherwise duplicate, the examination, and place a copy in the hands of each pupil. But there are numerous schools which lack duplicating machines. For teachers in these schools some other means for applying the test must be found. Any one of the following methods may be used. First, the entire test may be copied word for word by the pupils and then marked. This is tedious and time consuming. Second, the entire test may be written on the blackboard by the teacher. Each pupil could number a blank page of paper to correspond to the numbered statements, and then write *True* or *False* after the appropriate numbers. The only objection to this suggestion is the inconvenience of writing all the statements on the blackboard. Third,

the pupils may be asked to copy on blank paper, 1, 2, 3 and so on, according to the number of statements. The teacher can then read orally statement number 1 and instruct the pupils to make a check after the number 1 on their paper if the statement is true, but to make a cross if the statement is false. This is easily the most convenient way to give the examination. The chief objection to this last method is the difficulty some pupils have in apprehending statements presented orally, particularly if they are long and complicated. When the statement is presented visually the pupil has an opportunity to go back to it enough times to realize his possibility of understanding it. By one or another of these methods it is possible for any teacher anywhere to make use of this type of examination.

How should the *True-False* examination be scored? If a copy of the test has been placed in the hands of each pupil, the teacher can take an unused test sheet, fill it out correctly, lay the correct column of answers beside the pupil's column of answers, and quickly mark whether the pupil's answers are correct or incorrect. If a copy of the test has not been placed in the hands of each pupil, but each has instead written *True* or *False,* or made a check or cross after the number of each statement, the teacher can take a page of paper similar to that on which each pupil has indicated his answers, copy the numbers just as they are and just as they are spaced on each pupil's paper, write after each number the correct answer to the statement of that number, place this column of correct answers beside the column of pupil answers and mark those which are correct and incorrect. This last scoring method presupposes that pupils have used ruled paper, and that each has written his numbers in a vertical column according to a particular spacing recommended by the teacher. Last and best each pupil can score his own or his neighbor's paper. It is better for him to score his own.

If the method of pupil scoring is adopted, the teacher should read the correct answers while the pupil checks his own. If the pupil does not have a copy of the statements before him, the teacher should read each statement before giving the correct answer, in order that the pupil may know what statements he got correct or incorrect. When all the pupils' answers have been

marked and when all their scores have been computed and recorded on their examination paper, the teacher should ask all pupils who missed statement number 1 to hold up their hands, and then all pupils who missed number 2 to hold up their hands and so on. The teacher should make a record of the number of pupils missing each statement, and then collect all papers.

But why should this sort of examination be given at all? Wherein is it superior to the examination method in common use? In the first place, the *True-False* examination permits a teacher to cover a wider field of subject matter or a wider range of ability per unit of time. It may be made more representative of the total field of the pupils' study. In the case of the traditional examination the teacher is forced to select a very small number of questions. When we were students, almost as much of our ingenuity went into divining the kind of questions the teacher would ask as into reviewing. Now that we are teachers we have no reason to suppose that this practice has ceased. Rivalry among pupils in taking the traditional examination is largely a contest in guessing. The *True-False* examination can ask so many questions in a brief time that the only hope of the pupil is to study everything everyday and review everything when he reviews. Further, this extensive examining is fairer to pupils. The pupil has some basis for the time-honored excuse that he was so unlucky as to know everything except the particular questions asked by the teacher. The new type of examination rules this alibi out of court.

The use of this type of examination is likely to improve the relation between teacher and pupils. The traditional examination endangers a pleasant relationship because pupils more or less justly suspect that the score they make depends almost as much upon their conduct as upon their product. According to the pupils' suspicions the teacher uses her control over scores to vent a personal spite against the bad boy, and to express her appreciation of the "pet."

There is no question but what the pupils are partly correct. The traditional examination is so subjective that it is well-nigh impossible for the teacher to prevent her judgment from being influenced by fluctuations in her disposition or from being colored

by her affection for the pupil whose paper she is scoring. The new examination does away with these prolific sources of unpleasantness. It is so objective that the pupil can see for himself just how well he did. It permits self-scoring, and it convinces a pupil that the score he gets is the score he deserves.

The *True-False* examination is more enjoyable for the pupils. "Children cry for it" may be a bit exaggerated, but at any rate they hate it less. It offers an opportunity for a contest where the rules are fair, a chance for a larger degree of participation in the examination. It is agonizing for a pupil to describe at great length a knowledge which he does not possess in the hope that his command of English will camouflage his lack of information. Here is a question which was asked in a recent examination in educational measurement:

> Which three of the tests described by Whipple do you think would be of most service in an elementary school, if your school had a psychologist to apply them?

Consider the perspiration it must have cost a student to perpetrate this answer:

> The tests described by Whipple embraced most of the difficulties that would be embraced in problems of classroom instruction. I think his tests embrace a great variety of methods of approach and it seems difficult for me to think of just three to whom the presence of a psychologist in a school would give help. I would think it would be the tests in which knowledge of the workings of a child's mind and its growth and development would be most apparent since those not particularly trained might focus on others not of this kind. I fear it would be unwise to specifically mention just three when the number is so great which would fulfill all these requirements. Every teacher to be a psychologist would help all classroom measurement work of whatever kind greatly, I know since we cannot know of the influence of a test upon any group except by the mental reaction produced.

The *True-False* examination is also more enjoyable for the teacher because the scoring is easy, rapid and automatic when she does the scoring, and far more rapid when the pupils do the scoring. The pupils cannot well assist in scoring the traditional examination, and for the teacher to score forty verbose examination papers is time-consuming drudgery. Every moment of the time while scoring, the teacher must be profoundly concentrating upon what she is reading, for much of the time she must be sepa-

rating the chaff from the wheat where the chaff is cleverly painted to look like wheat. And along with this is a continual emotional strain caused by her resistance to the temptation to underscore some and overscore others.

The *True-False* examination is more educative to the pupils. The proposition that pupil scoring will relieve the teacher of much obnoxious drudgery, does not justify the inference frequently made that what is non-educative drudgery for the teacher will also be noneducative drudgery for the pupils. On the contrary we are of the opinion that the most favorable teaching opportunity that ever comes to a teacher is the period immediately following an examination. The pupil's interest to know what parts of the examination he missed and what he got correct is then at white heat. Witness the interested discussion among pupils immediately following an examination. It is inexcusable neglect of an educational opportunity not to capitalize these precious moments for correcting erroneous ideas, clinching right ideas, and filling up mental spaces where ideas are not. These values can best be realized by having each pupil score his own paper and by stopping to discuss points where pupils have trouble. Of course not every correct answer indicates knowledge, but the pupil himself usually knows when he knows. This examination is also more educative, because it is likely to be given more frequently. The experience of Kirby, Courtis, and others with practice tests shows that a pupil learns more during testing periods than during teaching periods. We really teach when we test. This examination covering as it can a wide range is an ideal method of review. It reveals to the pupils just where their difficulties lie. Testing is one of the best ways of teaching.

The *True-False* examination gives the teacher a fuller knowledge of conditions. The educative value of testing is so great that testing should be much more frequent than is now the case. If a method of testing is available which involves no drudgery to anyone, testing is likely to become more frequent, and this means more complete and timely information about the abilities and difficulties of the various pupils, and about the successes and failures of teaching efforts. We have already suggested that the teacher keep a record of the number or percent of pupils

missing each statement in the examination. This record will show what things have been well-learned or poorly learned, and well taught or poorly taught. Also it is a good thing for a teacher to check the effectiveness of her teaching. This can be done by finding the average of all pupils' scores and by comparing this average with the total number of statements in the examination or at least the total number of facts the teacher had really attempted to teach the pupils. If the average score is 20 out of a possible 40, the teaching efficiency is 50 percent.

Finally the *True-False* examination is a genuine honesty test, and shows the beginnings of a technic for measuring in satisfactory fashion this valuable character trait. Occasional and unannounced rescoring of each pupil's paper by his neighbor will catch the persistent cheat. It is better that he be discovered in school than in court. His discipline can usually be left to his fellow pupils, over whom he was attempting to gain an advantage dishonestly.

We have listed what appear to be the chief advantages of this type of examination or any other examination which is similarly objective. Most of these claims rest upon logical probability and a limited experience and not upon experimental data. This last is needed and will follow in time.

There are some limitations which have not yet been discussed. It is claimed first that this examination does not require the pupil to demonstrate a power to organize his materials. This is true in the sense that the pupil does not *describe in writing* a complicated mental organization but a statement can be so worded as to require an exceedingly complex mental organization before a correct answer can be unfailingly given. Consider the mental organization that must precede a correct answer to this simple statement: "If the trade winds blew east Peru would have luxuriant flora." If it is desired to test a pupil's power to word his thought a composition test may be given.

Again, it is claimed that this examination can test knowledge but not skill, knowledge but not the ability to do. Even skills, however, can be tested by this type of examination. To reason that trade winds blowing east would be warm, would absorb moisture from the Pacific, would become chilled in passing over

the Andes, would consequently deposit a heavy rainfall for Peru, which taken in conjunction with the equatorial climate would produce a luxuriant flora, is one sort of skill which this examination will test. Mathematical skills and the like which are too complicated to describe may be tested in at least two ways. An example or problem may be stated together with an answer. The pupil's task will be to determine by working the problem whether the answer given is true or false. Or instead, the teacher may work the problem on the blackboard for all the pupils and have them indicate whether her process was correct or incorrect.

Finally it is claimed that the teacher needs to know why a pupil is unable to answer the question about Peru and its flora. The *True-False* examination does not show just where the pupil's reasoning process went wrong or stopped altogether. It is not diagnostic. This criticism has some force. An examination should be as diagnostic as possible. If a teacher wished to know where the pupil's process broke down she could give a subsequent more detailed examination of this type. The statement: "The trade winds are warm winds" or "Warm winds have a larger capacity for water than cool winds," etc., would reveal whether the pupils were acquainted with the basic principles, facts and the like necessary to reason out the correct answer to: "If the trade winds blew east Peru would have luxuriant flora."

The traditional examination has certain advantages which will doubtless continue its existence. The *True-False* examination is a herald of newer and better types of examinations. But even now we have in the *True-False* examination one that may be used by any teacher anywhere to great advantage.